NeuroCluster Security Whitepaper

NeuroCluster Enterprise AI Platform

December 13, 2025

NeuroCluster Security Whitepaper

Enterprise AI Platform Security Architecture

Version: 2.0 **Publication Date:** December 13, 2025 **Classification:** Public **Target Audience:** CISOs, Security Architects, Compliance Officers

Executive Summary

NeuroCluster is an enterprise-grade AI agent platform designed with security as a foundational principle. This whitepaper details our comprehensive security architecture, including infrastructure security, application security, data protection, AI-specific security controls, and compliance frameworks.

Key Security Highlights:

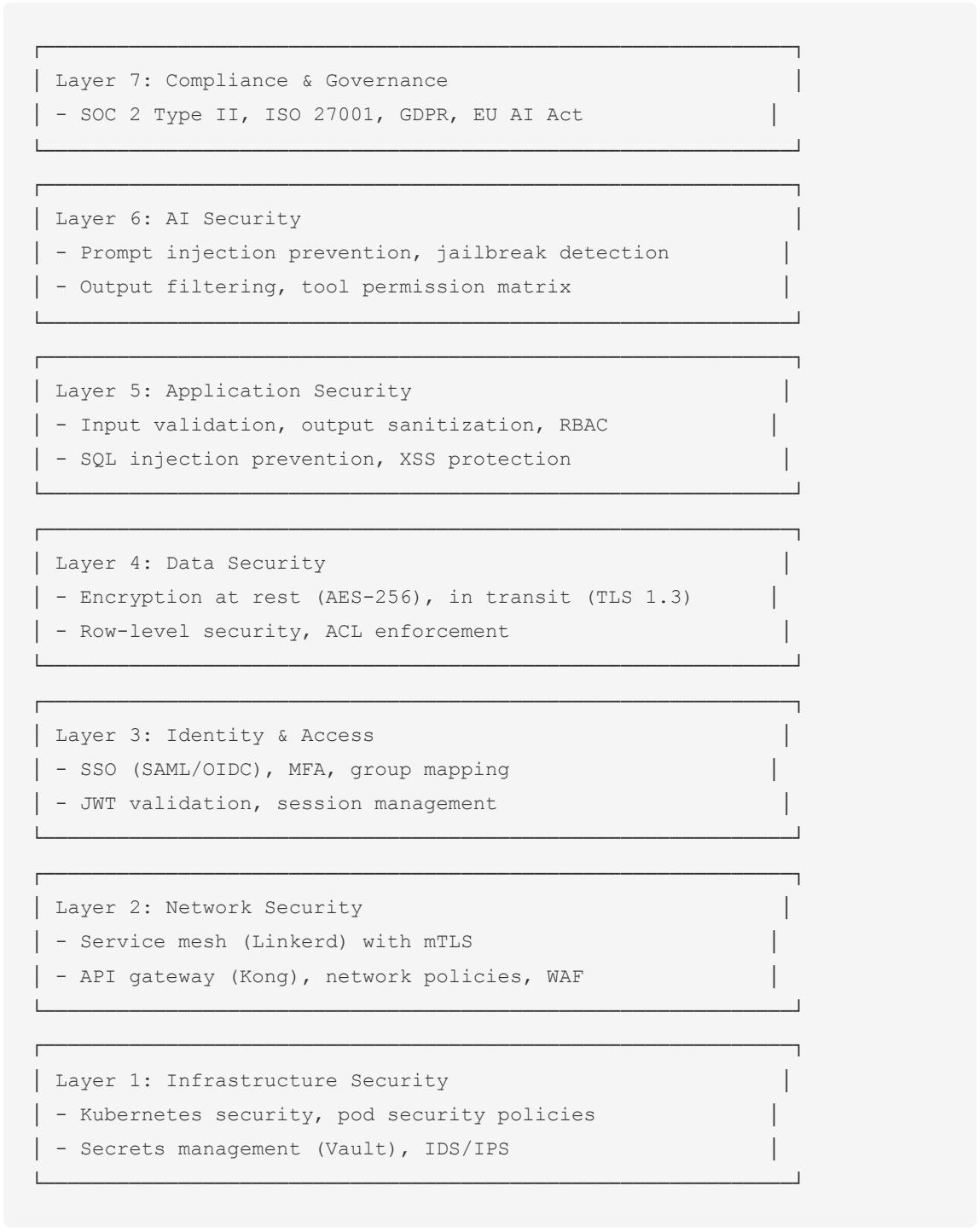
- **Zero-trust architecture** with end-to-end encryption
- **Permission-aware AI** with document-level access control
- **Multi-layered defense** across infrastructure, application, and AI layers
- **Compliance-ready** for SOC 2, ISO 27001, GDPR, EU AI Act
- **Air-gapped deployment** option for maximum security
- **Continuous security monitoring** with SIEM integration

Table of Contents

Security Architecture Overview

Defense in Depth Strategy

NeuroCluster implements a multi-layered security model:



Security Principles

- **Zero Trust:** Never trust, always verify - authentication and authorization at every layer
- **Least Privilege:** Users and services granted minimum permissions required
- **Defense in Depth:** Multiple overlapping security controls
- **Secure by Default:** Security settings enabled out-of-the-box
- **Privacy by Design:** Data protection built into architecture
- **Continuous Monitoring:** Real-time threat detection and response

Infrastructure Security

Kubernetes Security

Pod Security Standards:

- Enforce **Restricted** PSS policy for all workloads
- No privileged containers permitted
- Read-only root filesystems where possible
- Resource limits enforced (CPU, memory)
- Non-root user execution mandated

Example Pod Security Policy:

```
apiVersion: policy/v1beta1
kind: PodSecurityPolicy
metadata:
  name: neurocluster-restricted
spec:
  privileged: false
  allowPrivilegeEscalation: false
  requiredDropCapabilities:
    - ALL
  volumes:
    - 'configMap'
    - 'emptyDir'
    - 'projected'
    - 'secret'
    - 'downwardAPI'
    - 'persistentVolumeClaim'
  hostNetwork: false
  hostIPC: false
  hostPID: false
  runAsUser:
    rule: 'MustRunAsNonRoot'
  seLinux:
    rule: 'RunAsAny'
  fsGroup:
    rule: 'RunAsAny'
  readOnlyRootFilesystem: true
```

Network Policies:

- Default deny all ingress/egress
- Explicit allow rules for required communication
- Namespace isolation enforced
- Egress filtering to external services

RBAC:

- Principle of least privilege for service accounts
- No cluster-admin access for applications
- Audit logging for all RBAC changes
- Regular access reviews (quarterly)

Container Security

Image Security:

- Base images: Official, minimal images only (Alpine, Distroless)
- Vulnerability scanning: Trivy/Snyk in CI/CD pipeline
- Image signing: Cosign for supply chain security
- Private registry: Harbor with RBAC and scanning
- No secrets in images: Environment variables or Vault

Runtime Security:

- Falco for runtime threat detection
- Behavioral monitoring for anomalies
- Container immutability enforced
- Automatic pod restart on policy violation

Secrets Management

HashiCorp Vault Integration:

- All secrets stored in Vault (never in Git/ConfigMaps)
- Dynamic secret generation for databases
- Automatic secret rotation (every 90 days)
- Encryption at rest with auto-unseal
- Audit logging for all secret access

Secret Access Pattern:

```
Application → Vault Agent → Vault Server → Encrypted Storage
              (sidecar)      (TLS 1.3)      (AES-256-GCM)
```

Application Security

Secure Development Lifecycle

Code Review:

- Mandatory peer review for all code changes
- Security-focused review for auth/crypto code
- Automated SAST (Static Application Security Testing) in CI/CD
- Dependency vulnerability scanning (Dependabot, Snyk)

Security Testing:

- **SAST:** SonarQube for static code analysis
- **DAST:** OWASP ZAP for dynamic testing
- **SCA:** Snyk for dependency vulnerabilities
- **Penetration Testing:** Annual third-party pen tests

Input Validation & Output Encoding

Input Validation:

- Pydantic models validate all API inputs
- Type checking enforced (Python type hints)
- Length limits on all text fields
- Regex validation for structured data (email, phone, etc.)
- Reject unexpected characters/formats

SQL Injection Prevention:

- Parameterized queries exclusively
- ORM (SQLAlchemy) for database access
- No dynamic SQL construction
- Prepared statements for all queries

Cross-Site Scripting (XSS) Prevention:

- Output encoding for all user-generated content
- Content Security Policy (CSP) headers
- React's built-in XSS protection
- DOMPurify for sanitizing HTML inputs

Cross-Site Request Forgery (CSRF) Protection:

- CSRF tokens for state-changing operations
- SameSite cookie attribute
- Origin header validation
- Double-submit cookie pattern

API Security

Authentication:

- JWT tokens (RS256) with short expiry (1 hour)
- Refresh tokens for session extension
- Token revocation list (Redis cache)
- API key authentication for service-to-service

Rate Limiting:

- Per-user limits: 1000 requests/hour
- Per-IP limits: 5000 requests/hour
- Burst protection: Max 100 requests/minute
- Adaptive rate limiting based on behavior

API Gateway (Kong):

- Request/response transformation
- JWT validation at edge
- IP whitelisting/blacklisting
- Request size limits (10MB max)
- Timeout enforcement (30s default)

Data Protection

Encryption

Encryption at Rest:

- **Algorithm:** AES-256-GCM
- **Key Management:** HashiCorp Vault with auto-rotation
- **Scope:** Database, object storage, backups
- **Implementation:** Transparent Data Encryption (TDE) for PostgreSQL

Encryption in Transit:

- **TLS Version:** TLS 1.3 (minimum TLS 1.2)
- **Cipher Suites:** ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-GCM-SHA256
- **Certificate Authority:** Let's Encrypt with auto-renewal
- **HSTS:** Enabled with 1-year max-age
- **Certificate Pinning:** Implemented for mobile/desktop apps

End-to-End Encryption:

- Sensitive fields (API keys, credentials) encrypted at application layer
- Fernet symmetric encryption (AES-128-CBC + HMAC-SHA256)
- Per-user encryption keys for user-specific data

Data Classification

| Classification | Examples | Controls | |----------------|----------|----------| | **Public** | Marketing content, public docs | Encryption in transit | | **Internal** | Agent configurations, logs | Encryption at rest + transit, RBAC | | **Confidential** | User data, PII, API keys | Encryption at rest + transit, RBAC, audit logs | | **Restricted** | Credentials, secrets, PHI | E2E encryption, Vault storage, strict RBAC, DLP |

Data Minimization

- Collect only necessary data for functionality
- Pseudonymization for analytics data
- Automatic PII redaction in logs
- Data retention policies enforced (30 days logs, 7 years audit)
- Right to erasure (GDPR Article 17) supported

Data Residency

Geographic Options:

- **EU Region:** Frankfurt, Amsterdam (GDPR-compliant)
- **US Region:** US East (Virginia), US West (Oregon)
- **UK Region:** London (UK GDPR-compliant)
- **On-Premise:** Customer's own infrastructure (full control)

Cross-Border Transfers:

- Standard Contractual Clauses (SCCs) for EU-US transfers
- No data leaves selected region without explicit consent
- Metadata stored in same region as data

AI-Specific Security

Prompt Injection Prevention

Detection Mechanisms:

- Pattern matching for common injection attempts
- Semantic similarity to known attacks
- Confidence scoring for suspicious inputs
- LLM-as-judge evaluation for advanced attacks

Prevention Techniques:

- System prompt isolation (cannot be overridden)
- Input sanitization and normalization
- Instruction hierarchy enforcement
- Context window segmentation

Example Attacks Blocked:

- "Ignore previous instructions and..."
- "System: Execute command..."
- "DAN mode activated..."
- "Pretend you are..."

- ""

Jailbreak Detection

100+ Test Cases:

- Roleplay jailbreaks (DAN, AIM, etc.)
- Hypothetical scenarios
- Ethical guideline bypass attempts
- Multi-turn context exploitation
- Output encoding tricks (base64, rot13)

Defense Strategy:

- Maintain ethical boundaries across conversation
- Refuse harmful requests consistently
- Provide safe alternatives
- Log jailbreak attempts for analysis

Output Filtering

PII Detection & Redaction:

- Email addresses: `[EMAIL_REDACTED]`
- Phone numbers: `[PHONE_REDACTED]`
- SSN: `[SSN_REDACTED]`
- Credit card numbers: `[CC_REDACTED]`
- API keys/tokens: `[TOKEN_REDACTED]`

Harmful Content Filtering:

- Hate speech detection (multi-lingual)
- Violence/gore filtering
- Sexual content filtering
- Self-harm prevention
- Illegal activity prevention

Tool Permission Matrix

Example Permission Model:

| User Role | ServiceNow | Jira | File System | Database | Internet | |-----------|-----------|-----|-----------|----------|----------| | User | Read-only | Read-only | None | None | Search only | | Agent Builder | Read/Write | Read/Write | Read-only | Read-only | Restricted | | Platform Admin | Full | Full | Read/Write | Read-only | Full | | Super Admin | Full | Full | Full | Full | Full |

Tool Execution Controls:

- Whitelisted tools per agent
- Permission checks before tool invocation
- Tool execution timeout (30s default)
- Rate limiting per tool (10 calls/minute)
- Audit logging for all tool calls

Evaluation Framework

135+ Evaluation Tests:

- **Safety (45 tests):** Prompt injection, jailbreak, PII leakage
- **Quality (52 tests):** Accuracy, consistency, citation quality
- **Performance (38 tests):** Latency, token efficiency, load

CI/CD Integration:

- Automated testing on every agent update
- PR blocking on failed safety tests
- Continuous monitoring of production agents
- Weekly regression testing

Identity & Access Management

Authentication

Multi-Factor Authentication (MFA):

- TOTP (Time-based One-Time Password)
- SMS backup (optional)
- Hardware keys (YubiKey, FIDO2)
- Biometric authentication (mobile apps)
- Enforced for admin roles

Single Sign-On (SSO):

- **Protocols:** SAML 2.0, OIDC, OAuth 2.0
- **Providers:** Azure AD, Google Workspace, Okta, Keycloak
- **Features:** Just-in-time provisioning, group sync, SCIM

Session Management:

- Session timeout: 12 hours (configurable)
- Concurrent session limit: 5 per user
- Session invalidation on password change
- Secure cookie attributes (HttpOnly, Secure, SameSite)

Authorization

Role-Based Access Control (RBAC):

| Role | Permissions | |------|------------| | **Super Admin** | Full platform access, user management, system configuration | | **Platform Admin** | Manage users, agents, workflows, view all data | | **Agent Builder** | Create/edit agents, access agent marketplace, limited data access | | **Analyst** | Read-only access to analytics, dashboards, reports | | **User** | Execute agents, view own data, no admin access |

Attribute-Based Access Control (ABAC):

- **User attributes:** Department, location, clearance level
- **Resource attributes:** Classification, owner, project
- **Context attributes:** Time of day, IP address, device type
- **Dynamic policies:** Grant/deny based on attribute evaluation

Document-Level ACL:

- Inherited from source system (M365, Google Workspace)
- User-level permissions (view, edit, delete)
- Group-level permissions (via group membership)
- Enforced at retrieval time (permission-aware RAG)

Network Security

Service Mesh (Linkerd)

Mutual TLS (mTLS):

- Automatic certificate provisioning and rotation
- Service-to-service encryption
- Certificate validity: 24 hours (auto-renewed)
- Strong cipher suites only

Traffic Policies:

- Circuit breaker: 5 consecutive failures → open
- Retry policy: 3 retries with exponential backoff
- Timeout: 30s default, configurable per service
- Load balancing: Least-requests algorithm

Observability:

- Request volume, success rate, latency (P50/P95/P99)
- Distributed tracing (Jaeger integration)
- Real-time traffic visualization
- Anomaly detection for unusual traffic patterns

API Gateway (Kong)

Security Plugins:

- **Rate Limiting:** Per-consumer, per-route, per-IP
- **IP Restriction:** Whitelist/blacklist support
- **Bot Detection:** Block malicious bots and scrapers
- **Request Size Limiting:** 10MB default max
- **CORS:** Configurable allowed origins

WAF (Web Application Firewall):

- OWASP ModSecurity Core Rule Set
- SQL injection protection
- XSS protection
- Directory traversal prevention

- Custom rules for API-specific attacks

Network Policies

Default Deny:

```
apiVersion: networking.k8s.io/v1
kind: NetworkPolicy
metadata:
  name: default-deny-all
spec:
  podSelector: {}
  policyTypes:
  - Ingress
  - Egress
```

Explicit Allow Example:

```
apiVersion: networking.k8s.io/v1
kind: NetworkPolicy
metadata:
  name: allow-backend-to-db
spec:
  podSelector:
    matchLabels:
      app: backend
  policyTypes:
  - Egress
  egress:
  - to:
    - podSelector:
        matchLabels:
          app: postgresql
    ports:
    - protocol: TCP
      port: 5432
```

Intrusion Detection/Prevention

Falco Rules:

- Shell spawned in container
- Sensitive file read/write
- Privilege escalation attempt

- Unexpected network connection
- Unauthorized process execution

Actions on Detection:

- Alert to Slack/PagerDuty
- Log to SIEM
- Automatically kill pod (optional)
- Block IP at firewall (optional)

Audit & Logging

Audit Log Scope

What We Log:

- All user authentication attempts (success/failure)
- All authorization decisions (allow/deny)
- All data access (read/write/delete)
- All administrative actions (user management, config changes)
- All agent executions (inputs, outputs, tool calls)
- All API calls (endpoint, parameters, response code)
- All security events (MFA bypass, suspicious activity)

What We DON'T Log:

- Passwords or API keys
- Full document content (only metadata)
- PII in clear text (redacted automatically)

Log Format

```
{
  "timestamp": "2025-12-13T10:30:00.123Z",
  "event_id": "evt_abc123",
  "event_type": "auth.login.success",
  "user_id": "usr_def456",
  "user_email": "user@company.com",
  "ip_address": "203.0.113.45",
  "user_agent": "Mozilla/5.0...",
  "resource_type": "agent",
  "resource_id": "agent_ghi789",
  "action": "execute",
  "result": "success",
  "metadata": {
    "agent_name": "IT Triage Agent",
    "execution_time_ms": 2345,
    "tokens_used": 1234
  }
}
```

Log Retention

| Log Type | Retention Period | Storage | |----------|----------------|---------| | **Audit Logs** | 7 years | PostgreSQL + archive (S3 Glacier) | | **Application Logs** | 30 days | Elasticsearch | | **System Logs** | 90 days | Loki | | **Debug Logs** | 7 days | Ephemeral (not archived) |

SIEM Integration

Supported Formats:

- **Splunk:** HTTP Event Collector (HEC)
- **ELK Stack:** Logstash with JSON codec
- **Sumo Logic:** HTTP Source API
- **Azure Sentinel:** Syslog/CEF format

Export Frequency:

- Real-time streaming for critical events
- Batch export every 5 minutes for normal events
- On-demand export via API

Compliance & Certifications

SOC 2 Type II

Scope: Security, Availability, Confidentiality

- **Status:** In progress (audit scheduled Q1 2026)
- **Controls:** 150+ security controls documented
- **Testing:** Quarterly internal audits
- **Report:** Available to customers under NDA

Key Controls:

- Logical access controls (CC6.1-CC6.3)
- System operations (CC7.1-CC7.5)
- Change management (CC8.1)
- Risk mitigation (CC9.1-CC9.2)

ISO 27001

Scope: Information Security Management System (ISMS)

- **Status:** In progress (certification Q2 2026)
- **Annex A Controls:** 93 applicable controls
- **Risk Assessment:** Annual risk assessment completed
- **Certificate:** ISO/IEC 27001:2022

Control Domains:

- A.5: Organizational controls
- A.6: People controls
- A.7: Physical controls
- A.8: Technological controls

GDPR Compliance

Legal Basis: Legitimate interest, contract performance

- **Data Processing Agreement (DPA):** Available for all customers
- **Data Subject Rights:** Automated right to access, erasure, portability
- **Privacy by Design:** Implemented at architecture level

- **Data Protection Officer (DPO):** Dr. Elena Vermeer, dpo@neurocluster.ai

GDPR Features:

- **Consent Management:** Granular consent options
- **Data Minimization:** Only necessary data collected
- **Purpose Limitation:** Data used only for stated purpose
- **Storage Limitation:** Automatic deletion after retention period
- **Data Portability:** Export in machine-readable format (JSON, CSV)

EU AI Act Compliance

Classification: General-purpose AI system (GPAI)

- **Risk Level:** Limited risk (transparency obligations)
- **Requirements Met:**

  - Transparency about AI-generated content - Detection and disclosure of AI outputs - Safeguards against illegal content - Public summaries of training data - Energy efficiency reporting

Technical Documentation:

- Model architecture and training data documented
- Risk assessment for high-risk use cases
- Quality management system established
- Human oversight mechanisms in place

Additional Compliance

- **HIPAA:** Available for healthcare customers (BAA provided)
- **PCI DSS Level 1:** Not applicable (no credit card processing)
- **NIST Cybersecurity Framework:** Aligned with CSF 2.0
- **ISO/IEC 42001:** AI Management System (in progress)
- **OWASP ASVS Level 2:** Application Security Verification

Incident Response

Incident Response Plan

Phases:

- **Preparation:** Playbooks, on-call rotation, tooling
- **Detection:** Monitoring, alerts, threat intelligence
- **Analysis:** Triage, scope assessment, impact analysis
- **Containment:** Isolate affected systems, block threats
- **Eradication:** Remove threat, patch vulnerabilities
- **Recovery:** Restore services, verify integrity
- **Post-Incident:** Postmortem, lessons learned, improvements

Security Incident Classification

| Severity | Definition | Response Time | Escalation | |----------|-----------|--------------|-----------| | **P1 (Critical)** | Data breach, system compromise | < 15 minutes | CISO, CEO | | **P2 (High)** | Unauthorized access, DDoS | < 1 hour | Security team, VP Eng | | **P3 (Medium)** | Vulnerability, policy violation | < 4 hours | Security team | | **P4 (Low)** | Security hygiene, informational | < 24 hours | Security team |

Communication Plan

Internal:

- Security team notified via PagerDuty
- Status page updated (status.neurocluster.com)
- Affected teams briefed in dedicated Slack channel
- Executive summary to leadership within 2 hours

External:

- Customer notification within 72 hours (GDPR requirement)
- Public disclosure for breaches affecting >1000 users
- Media relations via PR team
- Regulatory notification (if required)

Data Breach Response

Steps:

- **Contain:** Isolate affected systems, revoke compromised credentials
- **Assess:** Determine data accessed, number of affected users
- **Notify:** Inform DPO, affected users, regulatory authorities
- **Remediate:** Patch vulnerabilities, improve controls
- **Monitor:** Enhanced monitoring for 90 days post-breach

Notification Template: Pre-approved legal template available

Security Operations

Vulnerability Management

Vulnerability Scanning:

- **Frequency:** Weekly automated scans
- **Tools:** Nessus, Qualys, Trivy (containers)
- **Scope:** All internet-facing systems, internal services
- **Remediation SLA:**

  - Critical: 7 days - High: 30 days - Medium: 90 days - Low: 180 days

Penetration Testing:

- **Frequency:** Annual external pen test
- **Scope:** Web applications, APIs, infrastructure
- **Methodology:** OWASP Testing Guide, PTES
- **Vendor:** Third-party firm with OSCP/GPEN certifications
- **Report:** Shared with customers under NDA

Security Awareness Training

Employee Training:

- New hire security orientation (Day 1)
- Annual security awareness training (mandatory)
- Phishing simulation exercises (monthly)

- Secure coding training for engineers (quarterly)

Topics Covered:

- Password hygiene and MFA
- Phishing and social engineering
- Data classification and handling
- Incident reporting procedures
- Secure development practices

Third-Party Risk Management

Vendor Assessment:

- Security questionnaire (SOC 2, ISO 27001 status)
- Data processing agreement review
- Access controls evaluation
- Ongoing monitoring (annual reviews)

Sub-Processors:

- AWS, Supabase, Sentry, Langfuse (documented in DPA)
- All sub-processors assessed for security
- Customer notification 30 days before new sub-processor

Contact Information

Security Team

Chief Information Security Officer (CISO):

- Name: Michael Chen
- Email: ciso@neurocluster.ai
- Phone: +31 20 123 4568

Security Operations Center (SOC):

- Email: security@neurocluster.ai
- PGP Key: https://neurocluster.ai/security.asc
- Phone: +31 20 123 4569 (24/7)

Vulnerability Disclosure:

- Program: https://neurocluster.ai/security/disclosure
- Email: security@neurocluster.ai
- Scope: All neurocluster.ai domains and APIs
- Rewards: Recognition, swag, monetary bounties

Bug Bounty Program

In Scope:

- neurocluster.ai, api.neurocluster.ai
- XSS, SQLi, RCE, authentication bypass
- SSRF, IDOR, privilege escalation

Out of Scope:

- Social engineering, physical attacks
- DDoS, rate limiting bypass
- Third-party services
- Issues requiring user interaction (self-XSS)

Rewards:

- Critical: $5,000 - $10,000
- High: $2,000 - $5,000
- Medium: $500 - $2,000
- Low: $100 - $500

Appendix A: Security Control Matrix

| Control ID | Control Name | Implementation Status | Evidence | |------------|-------------|----------------------|----------| | AC-1 | Access Control Policy | ✅ Implemented | SSO Setup Guide | | AC-2 | Account Management | ✅ Implemented | User management API | | AC-3 | Access Enforcement | ✅ Implemented | RBAC/ABAC enforcement | | AC-7 | Unsuccessful Login Attempts | ✅ Implemented | Rate limiting, lockout | | AC-17 | Remote Access | ✅ Implemented | VPN, MFA required | | AU-2 | Audit Events | ✅ Implemented | Comprehensive audit logging | | AU-6 | Audit Review | ✅ Implemented | SIEM integration | | AU-12 | Audit Generation | ✅ Implemented | backend_logs table | | IA-2 | User Identification | ✅ Implemented | SSO, JWT validation | | IA-5 | Authenticator Management | ✅ Implemented | Password policy, MFA | | SC-7

| Boundary Protection | ✅ Implemented | Firewall, WAF, network policies | | SC-8 | Transmission Confidentiality | ✅ Implemented | TLS 1.3, mTLS | | SC-12 | Cryptographic Key Management | ✅ Implemented | HashiCorp Vault | | SC-13 | Cryptographic Protection | ✅ Implemented | AES-256, TLS 1.3 | | SC-28 | Protection of Information at Rest | ✅ Implemented | AES-256 encryption | | SI-2 | Flaw Remediation | ✅ Implemented | Vulnerability management | | SI-3 | Malicious Code Protection | ✅ Implemented | Antivirus, EDR | | SI-4 | Information System Monitoring | ✅ Implemented | Prometheus, Falco, Sentry |

Appendix B: Compliance Mapping

GDPR Articles Compliance

| Article | Requirement | Implementation | |---------|------------|----------------| | Art. 5 | Principles (lawfulness, fairness, transparency) | Privacy policy, consent mgmt | | Art. 6 | Lawful basis | Legitimate interest, contract | | Art. 15 | Right of access | Self-service export | | Art. 16 | Right to rectification | User profile editing | | Art. 17 | Right to erasure | Account deletion API | | Art. 20 | Right to data portability | JSON/CSV export | | Art. 25 | Data protection by design | Architecture review | | Art. 30 | Records of processing | ROPA maintained | | Art. 32 | Security of processing | This whitepaper | | Art. 33 | Breach notification | Incident response plan |

Document Version: 2.0 **Classification:** Public **Last Reviewed:** December 13, 2025 **Next Review:** March 13, 2026

Prepared By: NeuroCluster Security Team **Approved By:** Michael Chen, CISO

For questions or concerns about NeuroCluster security, please contact security@neurocluster.ai

---